Synopsis

Epidemiological profiling of population-specific risk factors and validation of novel genetic variants and deep learning-driven mitosis detection in breast cancer patients

Candidate
K. Suganya
20PHBTF001
(Ph.D. Biotechnology)

Supervisor
Dr. S. Sumathi
Associate Professor
Department of Biochemistry, Biotechnology and Bioinformatics

Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore - 641 043

February, 2024

Introduction

Breast cancer is the most prevalent cancer among women and a leading cause of cancer-related mortality worldwide. Over 2.3 million new cases and 685,000 deaths from breast cancer occurred in 2020. Projections indicate that by 2030, the global number of new breast cancer cases may reach 2.7 million annually, with 0.87 million deaths. By 2040, the burden of breast cancer is predicted to increase to over 3 million new cases and 1 million deaths every year because of population growth (Arnold *et al*., 2022). According to the National Cancer Registry Programme, India, breast cancer is the leading cause of cancer among Indian women, and more than 50% of Indian women are diagnosed in advanced stages 3 and 4. In Tamil Nadu the rate of new cases increased from 42.9 per 1,00,000 women in 2013 to 49.8 per 1,00,000 women in 2021 (Sathishkumar *et al*., 2022). The reason for the low breast cancer survival rate among Indian women is a lack of awareness and poor early screening and diagnosis. While breast cancer is more common in women, it can also affect men, with much lower frequency, accounting for 1 in 100 diagnosed cases and less than 1% of all male cancer cases (Campos *et al*., 2021).

Risk factors for breast cancer are broadly categorized into non-modifiable intrinsic factors and modifiable external factors. Intrinsic factors, determined by genetics and biology, include age, sex, race, exposure to endogenous steroid hormones, benign proliferative breast lesions, and genetic susceptibility (Rajaneesh, 2021). Modifiable external factors, influenced by lifestyle choices, encompass physical inactivity, alcohol consumption, smoking, pregnancies, dietary habits, and hormonal interventions. Understanding these factors is crucial for effective prevention and intervention strategies (Taha and Eltom, 2018).

Epidemiological studies are vital in identifying and understanding modifiable risk factors, providing a foundation for preventive strategies. Focusing on these factors allows the design of interventions and public health initiatives to reduce breast cancer incidence (Iacoviello *et al*., 2021). This approach emphasizes the importance of lifestyle-based prevention strategies to combat this prevalent disease and improve public health outcomes (Osei-Afriyie *et al*., 2021).

The past two decades of advancements in Next-generation sequencing (NGS) are pivotal in genomics, particularly in diagnosis, prognosis, and treatment. Various NGS platforms offer multiple approaches, including whole-genome sequencing, whole-exome sequencing (WES), and targeted sequencing (Conway *et al*., 2019). WES is particularly significant in medical genetic research, efficiently covering nearly all protein-coding regions within the human genome. Despite WES's ability to identify variants, validation through

Sanger sequencing remains crucial, ensuring comprehensive and precise genetic analysis in clinical genomics. The combined use of WES and Sanger sequencing contributes significantly to advances in genomics and personalized medicine (Qin, 2019).

Sequencing helps to identify the genes and variants in the development and progression of cancer. In the context of cancer, tumor cells display distinct characteristics or hallmarks that set them apart from normal cells. These hallmarks include uncontrolled proliferation, genomic instability, and the ability to evade apoptosis (Huang *et al*., 2021). The cancer cells undergo specific modifications in various cell signalling pathways, which include estrogen receptor (ER), HER2/Neu (ERBB2) PI3K/AKT/mTOR Pathway, *TP53*, *BRCA1* and *BRCA2*, MAPK/ERK, Wnt/β-Catenin, Notch signalling, TGF-β (Transforming Growth Factor-beta), NF-κB (Nuclear Factor-kappa B) and hedgehog pathway. These pathways are essential for transmitting signals within cells and coordinating their activities (Sahu and Pattanayak, 2020). The pathways are altered by various risk factors, leading to the uncontrolled growth and survival of tumor cells. These alterations can be attributed to several factors, including mutations in oncogenes and tumor suppressor genes. Understanding these molecular mechanisms is crucial in cancer research and treatment. It provides insights into the development of targeted therapies to mitigate these alterations, ultimately contributing to more effective cancer treatments and improved patient outcomes (Glaviano *et al*., 2023).

Breast cancer grading is a crucial component of the diagnostic process, providing valuable information about the characteristics of the cancer cells and their potential for aggressiveness. The three main features in breast cancer grading are tubule formation, nuclear pleomorphism, and mitotic count. Within the realm of diagnosis, the mitotic cell count emerges as a crucial biomarker for assessing the aggressiveness and prognosis of breast cancer. Traditionally, pathologists have manually scrutinized histopathological images under high-resolution microscopes to identify mitotic cells. However, this process is laborious, time-consuming, and subjective, primarily due to the difficulty of distinguishing between mitotic and normal cells (Ibrahim *et al*., 2022). Artificial intelligence-based (AI-based) techniques have been developed to detect mitotic cells within histopathology images. AI algorithms can analyze large volumes of histopathology images much faster than human observers. Deep learning techniques have recently emerged as a viable solution, offering outstanding performance while minimizing computational costs (Pan *et al*., 2021). Integrating AI technologies into clinical practice can enhance breast cancer diagnosis efficiency and precision, ultimately improving patient outcomes (Pantanowitz *et al*., 2020).

The primary goal of personalized breast cancer medicine is to provide the most effective treatment by categorizing subtypes and tailoring regimens based on patient characteristics and therapy response. Specific therapeutic strategies target individual molecules or mutated genes to identify genetic variations driving selective cell growth in tumors. The lack of safe drugs poses a challenge to disease control. Biomarkers, therapeutic targets, and data mining play crucial roles, assisting clinicians in selecting effective regimens to overcome resistance. Gene mutation detection has significantly impacted disease management, emphasizing a personalized approach focused on tumor-specific molecular alterations (Jeibouei *et al*., 2019).

Breast cancer exhibits diverse risk factors and genetic alterations across different populations worldwide. Addressing the population-specific risk factors and rare genetic variants is challenging, and it is crucial for developing targeted therapy and reducing the impact of breast cancer on affected individuals. The existing knowledge on breast cancer risk factors and the role of rare genetic alterations in disease progression is not fully explored in the Tamil Nadu population. Bridging the gap between risk factors, molecular insights, and rare population-specific variants will lead to more effective and targeted therapeutic interventions.

With this background, the current study aimed to evaluate the epidemiological risk factors and genetic factors influencing disease progression in breast cancer patients in specific populations. A hospital-based cohort study was performed to identify the patient's risk factors. We wanted to know the prevalence of rare variants in representative samples among the selected population. So, we did whole exome sequencing to identify the rare variants on fresh tumor tissue samples obtained from six patients diagnosed with primary breast cancer. The six samples were selected based on the different breast cancer subtypes, and the variants were validated by Sanger sequencing. Furthermore, we utilized deep learning to detect mitotic figures in histopathology images, which enabled us to assess the rate of breast cancer proliferation due to the rare variants. This comprehensive study involves epidemiological, genetic, and histopathological analyses, providing insights into the complex factors underlying breast cancer progression.

Hypothesis

The hypotheses for the present study are as follows:

Null hypothesis

- There is no significant association between a specific population's epidemiology profile and breast cancer incidence.

- There are no novel genetic variants associated with breast cancer in the specific population.

Alternate hypothesis

- There is a significant association between the epidemiology profile of a specific population and breast cancer incidence.
- There are novel genetic variants associated with breast cancer in a specific population.

Objectives

The objectives framed for the present investigation were as follows:

- To assess the epidemiological profile of breast cancer-associated risk factors.
- Identifying rare genetic variants associated with breast cancer by whole exome analysis.
- Validation of the identified novel genetic variants using Sanger sequencing.
- To detect mitosis and evaluate breast cancer progression using histopathology images by deep learning.

Phase I

In the study's first phase, detailed information was collected from patients visiting the oncology division of Ramakrishna Hospital, Coimbatore. Ethical clearance was obtained from Ramakrishna Hospital, Coimbatore (EC/2019/0411/CR/57) and Avinashilingam University, Coimbatore (AUW/IHEC/BT-22-23/FHP-02). Informed consent was obtained from the patients after the study was explained clearly in their understandable bilingual language. Inclusion criteria of the study were women aged > 20 years who were diagnosed with breast cancer. Exclusion criteria for the study were women who did not have breast cancer or any other cancer. Demographic, reproductive, lifestyle, and environmental factors, clinical characteristics, and family history of breast cancer were collected through a structured questionnaire. The sample size was determined using the Raosoft calculator. The data was statistically analyzed with SPSS Inc. version 20, which involved descriptive analysis, Pearson chi-square tests for variable frequency distribution, and a 95% confidence interval to assess the association between risk factors and breast cancer. A significance level of $p < 0.05$ was considered statistically significant.

The study cohort included both primary breast cancer patients and those undergoing cancer treatments. Initially intended for 377 individuals, the study exceeded expectations with 517 respondents, resulting in a larger dataset that promises to yield more robust insights. Demographic factors (age, gender, race/ethnicity, family history) are critical in breast cancer risk assessment, guiding targeted interventions for early detection and timely treatment.

Among individuals aged 21 to 30, breast cancer was seen in 4.45% (95% CI: 0.02 - 0.05) of the population. The 31-40 age group constituted 25.92% (95% CI: 0.22 - 0.30) of cases. The highest proportion was observed in the 41-50 age group, with 35.40% (95% CI: 0.31 - 0.39) of patients. For those aged 51-60, it represented 20.50% (95% CI: 0.17 - 0.24) of cases. In the 61-70 age group, breast cancer cases were at 11.60% (95% CI: 0.09 - 0.15). The lowest was observed in the age group 71-80 at 2.13% (95% CI: 0.01 - 0.03). The average age of the study population was 47.40 years. Notably, among the 517 patients, breast cancer was predominantly found in females, accounting for 99.81% of cases (95% CI: 0.98 - 0.99), with males constituting only 0.19% (95% CI: 0.001 - 0.005). Marital status showed that 98.45% of married individuals were affected (95% CI: 0.97 - 0.99) compared to 1.55% of unmarried individuals (95% CI: 0.01 - 0.03). Education level exhibited a strong association, with illiterate individuals at 24.76% (95% CI: 0.21 - 0.29), primary school completers at 26.50% (95% CI: 0.22 - 0.30), secondary and higher secondary education individuals at 29.60% (95% CI: 0.01 - 0.03), and graduates at 19.14% (95% CI: 0.16 - 0.22) showing varying incidence rates. However, employment status did not display statistical significance, with 34.43% of breast cancer cases in employed individuals (95% CI: 0.30 - 0.38) and 65.57% in unemployed individuals (95% CI: 0.61 - 0.69). These findings provide crucial insights into the demographic characteristics that were prominent among breast cancer patients.

Reproductive factors such as menarche, menopause, pregnancy, and breastfeeding play a significant role in causing breast cancer. Age at menarche showed a potential association with breast cancer, where the risk increased with a later onset. For those who experienced menarche at age 13, 7.17% were affected (95% CI: 0.05 - 0.09), while those at age 16 had a higher prevalence of 43.22% (95% CI: 0.39 - 0.48). The p-value of 0.781 and a chi-square (χ^2) value of 133.48 showed that there was no statistically significant association between age of menarche and breast cancer. Regarding age at first full-term pregnancy, the chi-square (χ^2) value was found to be 100.24, and a p-value of 0.418, and childbirth before age 30 was linked with reduced risk of breast cancer. In the case of women whose childbearing age was before 30, approximately 81.39% were affected (95% CI: 0.78 - 0.85) compared to those age of delivery at or after 30 had a prevalence of 18.61% (95% CI: 0.15 - 0.22). A p-value of 0.406 and a chi-square (χ^2) value of 100.67 revealed no significant implications of breast cancer associated with age at first childbirth. Breastfeeding duration did not exhibit a significant association with breast cancer risk, with percentages ranging from 0.19% to 46.14% and a χ^2 value of 241.00 (p-value: 0.560). Similarly, abortions did not show a significant link with breast cancer, as indicated by a χ^2 value of 49.62 (p-value:

0.448). However, menopausal status demonstrated a substantial influence on breast cancer risk. Premenopausal women have a prevalence of 56.01% (95% CI: 0.52 - 0.60), whereas postmenopausal individuals had 43.99% affected (95% CI: 0.40 - 0.48). The χ^2 value of 502.64 (p-value: 0.000) signified a significant association between menopausal status and breast cancer risk. Women experiencing menopause before the age of 50 faced a notably higher risk of 78.41% (95% CI: 0.73 - 0.84), while those at or after 50 exhibited a lower risk of 21.59% (95% CI: 0.16 - 0.27) with strong association of χ^2 value 788.46 and p-value 0.000. These findings provide the substantial impact of menopausal status on breast cancer risk.

Lifestyle factors such as behaviors, habits, and diets of the individuals in their day-to-day activities can influence their probability of developing breast cancer. Considering dietary habits, 68.28% of the surveyed population consumed fresh fruits and vegetables daily (95% CI: 0.64 - 0.72), and 83.37% (95% CI: 0.80 - 0.87) consumed meat frequently. Notably, 23.79% were engaged in daily exercises (95% CI: 0.20 - 0.27), while 53.97% exercised rarely or never (95% CI: 0.50 - 0.58). A significant association emerged between exercise frequency and breast cancer, with a χ^2 value of 75.93 and a p-value of 0.008. Less physical activity regulates oxidative stress in the body, which may trigger the development of cancer. Water intake displayed a substantial connection to breast cancer, with a χ^2 value of 122.72 and a p-value of 0.046 showing significant relationships. Dehydration can affect hormone production and function, and disruptions in hormonal balance may contribute to cancer development. Sleep duration exhibited statistical significance, with χ^2 value of 193.52 and a p-value of 0.006. Insufficient sleep disturbs the circadian rhythm, the biological clock, and may lead to a progression of cancer. 19.54% (95% CI: 0.16 - 0.23) cases were exposed to pesticides and other exposures, including cement work, showed a 6.58% association (95% CI: 0.05 - 0.09). It showed that exposure to pesticides and cement may be one of the prominent factors causing breast cancer. These results highlight the importance of lifestyle changes, promoting regular exercise, sufficient hydration, and adequate sleep to lower the chance of developing breast cancer.

Examining the clinical features of breast cancer patients helps in treatment adherence, side effect management, and overall health. There was a significant association ($\chi^2 = 78.83$, $p = 0.004$) between comorbidities and breast cancer. Notably, coexisting conditions such as high blood pressure (14.51%, 95% CI: 0.11 - 0.18), diabetes (11.80%, 95% CI: 0.09 - 0.15), and combinations of diabetes with other conditions showed substantial prevalence. The side of the primary breast tumor displayed a highly significant association ($\chi^2 = 184.45$, $p = 0.000$)

with breast cancer incidence, with 41.39% on the left side and 55.13% on the right and 3.48% on both sides of the breast. A highly significant relationship ($\chi^2 = 219.06$, $p = 0.000$) was observed between treatment modalities and breast cancer patients, with surgery (27.85%, 95% CI: 0.24 - 0.32), chemotherapy (46.03%, 95% CI: 0.42 - 0.50), radiotherapy (10.44%, 95% CI: 0.08 - 0.13), and combinations being prominent. There was no significant correlation ($\chi^2 = 286.47$, $p = 0.612$) between the reported side effects and breast cancer treatment; hair loss was a commonly reported side effect. Various surgical procedures had different prevalence rates, but there was no significant relationship ($\chi^2 = 299.57$, $p = 0.399$) observed between surgical history and breast cancer. Clinical characteristics of breast cancer patients revealed substantial associations between comorbidities, tumor location, and treatment, indicating the importance of personalized treatment strategies based on individual patient traits.

Evaluating familial history in breast cancer patients is essential for assessing genetic factors and overall risk. The survey examined the prevalence of breast cancer in first-degree relatives of individuals affected with breast cancer. Around 0.58% of breast cancer patients (95% CI: 0.001 - 0.013) had a mother previously diagnosed with breast cancer, while about 0.39% (95% CI: 0.002 - 0.008) had a sister with a prior diagnosis of breast cancer. However, a substantial majority, 99.03% (95% CI: 0.98 - 0.99), had no first degree of breast cancer. Statistical analysis with a chi-square (χ^2) test of 107.41 and p-value of 0.242 indicated no significant association between first-degree relatives' breast cancer history and the patients. Notably, no cases of breast cancer were reported among second-degree relatives. Examining major inheritable diseases, approximately 60.35% (95% CI: 0.56 - 0.64) of the study population had a family history of diabetes. In contrast, 39.65% (95% CI: 0.35 - 0.45) had no family history of such diseases. A p-value of 0.821 and a chi-square (χ^2) value of 39.85 represent the statistical analysis did not provide strong evidence supporting a significant association between major inheritable diseases in the family and individuals' risk of developing breast cancer.

The outcome of phase I highlights the importance of a holistic approach to risk assessment and prevention of breast cancer. Factors including lifestyle choices, reproductive history, and comorbid conditions play a role in an individual's susceptibility to breast cancer. These findings emphasize the need for personalized risk assessment and tailored preventive strategies to minimize the impact of breast cancer on people and systems of health care. Moreover, the research highlights the importance of early detection and access to healthcare, particularly for individuals facing socioeconomic challenges.

Phase II

Once we identified the risk factors, we did a whole exome analysis to determine the disease-causing variants and associated diseases in selected breast cancer patients in phase II. The genomic information from exome sequencing was then analyzed with a bioinformatics approach to identify rare genetic variations in the specific population. By examining the rare genetic alterations in breast cancer patients, we can understand the molecular mechanisms underlying the disease and identify potential targets for personalized treatments in the specific population. The rare variants tend to be population-specific, which is the main reason for disease development in the specific population compared to the common mutation in breast cancer. The findings from phase II, when combined with the comprehensive epidemiological data from phase I, contribute to a more holistic understanding of the complex interplay between genetic, environmental, and lifestyle factors in breast cancer etiology.

Tumor tissue samples were collected from 6 primary breast cancer patients (BC-1 to BC-6) (5-Female and 1-Male) and stored in phosphate-buffered saline at -20°C for DNA isolation. DNA was isolated from tumor tissue using QIAamp DNA Mini Kit (CA, USA). The quality of the isolated DNA was also checked using gel electrophoresis. DNA quantification was performed using Nanodrop (NanoDrop™ 1000 Spectrophotometer, Thermofisher). The absorbance ratio at 260 nm to 280 nm (A260/A280) for purity check of the DNA sample was done. The WES was done at Neuberg Supratech, Ahmedabad. The sequencing was performed using paired-end (PE) reads with a read length of 150 base pairs. The platform used for sequencing was the Illumina NovaSeq 6000. Additionally, the sequencing depth covered per sample was >8GB. The identified variant genes were analyzed using the FunRich software tool (http://www.funrich.org).

WES is done in the breast tumor tissue removed during mastectomy. Exome profiling may lead to the identification of new biomarkers associated with breast cancer subtypes or progression. The ages of six patients were found to range between 37 to 74. Four patients had tumors located on the right side, while two patients had tumors located on the left side. A family history of cancer was noted in two patients. All patients were diagnosed with DCIS (Ductal Carcinoma In Situ) and had Grade III tumors, suggesting a high level of cell abnormality and aggressiveness. All individuals presented a T2 tumor stage, indicating a tumor size ranging from more than 2 cm and not more than 5 cm. BC-2 patient alone displayed a T1c tumor stage, suggesting a smaller tumor size of 2 cm or less. Metastasis status varied, with some patients having micrometastases and metastasis and others having no distant metastasis. Patients had varying receptor statuses of ER, PR, and HER2, which can

influence treatment decisions. Ki-67 proliferation index also varied among patients, with some patients showing high levels of divisions.

The concentration and purity of the isolated DNA were assessed. The quality of the quantified DNA was confirmed. The good quality of the isolated DNA suggests that the DNA fragments are intact and do not appear degraded or fragmented. DNA fragments were prepared and tagged with specific adapters to enable sequencing during library preparation. The insert size analysis of the library was performed using the Tapestation 4150 (Agilent), which was equipped with high-sensitive D1000 screen tapes, confirming that the average fragment size for all patient libraries fell within the desired range of 350–450 base pairs. Specifically, the insert sizes for patients BC-1 to BC-6 were determined as follows: 451 bp, 335 bp, 382 bp, 356 bp, 357 bp, and 419 bp. This quality control step ensured the libraries were prepared with consistent and appropriate insert sizes.

Sequencing quality control involves verifying the quality of the sequencing run. The Phred quality score (Q-score) is a logarithmic scale commonly used in DNA sequencing to represent the accuracy of base calls in sequencing reads. The Phred score of 36 was maintained throughout all base pairs in the sequencing read of six breast cancer patients. It corresponds to a base call accuracy of 99.9%, meaning there is a 1 in 1,000 chance of an incorrect base call at that position. It indicates that the sequencing data is of very high quality and is associated with a low error rate. Specifically, a total of 1,01,213 variants were initially detected across all the breast cancer patients in the study. After applying filters to annotate the variants, 857 rare variants were identified.

Among the 857 significant variants identified, all were found to be heterozygous. Heterozygous variants imply that one of the two copies of a gene is mutated, indicating a potential role of these genetic changes in the development or progression of breast cancer. Additionally, all of these variants were located in the exonic region of the genes, which are the coding regions responsible for producing proteins. Any mutations occurring in these coding regions have the potential to impact the structure and function of the proteins and may contribute to abnormal cell growth or other hallmarks of cancer. Furthermore, all of these variants were classified as nonsynonymous single nucleotide variants (SNVs), resulting in amino acid changes in the protein sequence and potentially affecting protein structure and function. In the context of breast cancer, these changes could influence signalling pathways, cell cycle regulation, and DNA repair mechanisms. In the case of BC-4, two specific gene mutations were identified. One of the mutations in BC-4 occurred as a synonymous single nucleotide variant (SNV) in the gene *SCYL1* in exon 16 (c.2148C>T and c.2199C>T),

resulting in genetic changes in the DNA sequence impact gene expression, potentially influencing the cellular environment related to breast cancer. Another mutation in BC-4 was identified as a stop-gain mutation that occurred in exon 25 of the *ULK4* gene (c.1678C> 7 and c.2584C>T), which might lead to the premature termination of protein synthesis during translation. The loss of proper protein function can disrupt normal cellular processes, contributing to the uncontrolled growth characteristic of cancer cells.

The most common comorbidity observed in breast cancer patients was cardiomyopathy. The molecular pathways implicated in cardiomyopathy, including abnormal calcium handling and signalling, oxidative stress, and fibrosis, may intersect with pathways involved in breast cancer progression. Few disease conditions and environmental factors that elevate inflammation and oxidative stress levels can serve as risk factors for both breast cancer and cardiovascular disease. Robust epidemiological evidence indicates that reduced physical activity, dietary patterns, environmental exposures, and clinical conditions like diabetes and family history are linked with an elevated risk of both cardiovascular disease and breast cancer. The common gene identified among all breast cancer patients is *TTN* (titin). *TTN* encodes titin, the most significant known protein, which plays a crucial role in muscle contraction, including cardiac muscle. Mutations in the *TTN* gene have been associated with various heart-related conditions, including cardiomyopathies. These results suggest that individuals might carry specific genetic variants that predispose them to both breast cancer and cardiomyopathy. These variants could affect genes or pathways involved in the progression of both conditions. If a person with breast cancer develops cardiomyopathy, their treatment plan may need to be adjusted to minimize the risk of further cardiac complications. One significant finding emphasizes the importance of thorough cardiac evaluation before initiating breast cancer treatment.

In our study of breast cancer patients, we identified variants in coding regions of the genes after filtering. However, we want to explore the potential involvement of non-coding regions, specifically introns, in breast cancer progression. Mutations in introns do not directly impact protein sequences. But introns are large segments of DNA that make up a significant portion of human genes and are essential in gene expression. In the exome analysis of breast cancer, we applied a specific criterion, considering only intronic variants with an allele frequency of <0.05. The results showed a total of 183 intronic variants in six breast cancer patients.

We also looked for mutations in *BRCA1* and *BRCA2* genes, recognized as hotspot genes frequently mutated in various cancers, including breast cancer. In our analysis, all

breast cancer patients except BC-2 showed mutations, and variants in the *BRCA* genes were already reported in the dbSNP database. Identifying mutations in these genes is vital for assessing the risk of breast cancer, implementing preventive measures, and developing targeted treatment strategies for individuals with an elevated genetic predisposition to breast cancer. This underscores the significance of gene analysis in informing personalized approaches to breast cancer risk and management.

Functional enrichment analysis of candidate genes involves examining the biological functions, pathways, and processes significantly associated with the genes identified as candidates in the particular study. Functional enrichment analysis showed that the genes are involved in the primary biological process, including signal transduction, G-protein coupled receptor signalling pathway, positive regulation of gene expression, synapse assembly, and positive regulation of transcription from RNA polymerase II that holds significant relevance in breast cancer. Signal transduction pathways are crucial in breast cancer. Dysregulation of these pathways can lead to uncontrolled cell growth and proliferation, which are hallmark features of cancer cells. Aberrant signalling can contribute to breast cancer development primarily through growth factor receptors.

The association of candidate genes with cellular processes at the plasma membrane, cytosol, and cytoplasm underscores their involvement in breast cancer. Alterations in the plasma membrane can impact cell signalling. Changes in membrane receptors and transporters can affect the response to external stimuli, including growth factors and hormones that influence breast cancer cells. Cytosol and cytoplasm are the cellular compartments central to various cellular processes, including signal transduction and regulating intracellular pathways.

Dysregulation of signal transduction pathways, aberrant plasma membrane receptor activity, and altered protein binding events are interconnected processes that contribute to the oncogenic properties of breast cancer cells. These events can lead to uncontrolled growth, survival, and invasion, all of which are hallmarks of cancer. In summary, the candidate genes identified in the study and their associated biological processes and molecular functions provide insights into the molecular mechanisms underlying breast cancer. Dysregulation in these processes can drive the development and progression of breast cancer, making them important targets for further research and potential therapeutic interventions.

Among the significant variants identified, three were novel variants that have not been previously reported in breast cancer. These new variants were identified in the genes *MRPL13, MYPBC3,* and *PTCH1*. The *PTCH1* gene variant was identified in two different

breast cancer patient samples of BC-4 and BC-6. On further verification using the VarSome database, additional information regarding these variants and their associations with other cancer types was analyzed. *PTCH1* variant was found to be reported in different cancer types, such as endometrial cancer and colorectal cancer. The *MYBPC3* variant plays a significant role in several cancer types, including esophagogastric cancer, head and neck cancer, endometrial cancer, and myeloproliferative neoplasms. The presence of this specific variant in breast cancer patients highlights that genetic alteration may be population-specific due to epidemiological risk factors.

This study contributes significantly to our understanding of breast cancer at the genomic level, unravelling novel variants, elucidating potential risk factors, and providing a foundation for personalized therapeutic interventions. Integrating genomics into clinical practice holds promise for advancing precision medicine in oncology.

Phase III

The biological consequences of the specific genetic alteration identified from phase II results were validated in phase III, and its potential impact on the development and or progression of the disease was assessed. By characterizing the functional consequences of these genetic variations, their potential role in disease development or progression was determined. This phase provided critical insights into the mechanisms through which specific genetic alterations contribute to the pathophysiology of breast cancer within the selected population.

The specific primers were designed to the sequence of the novel variant, enabling efficient and accurate amplification and sequencing. Touchdown PCR was done to amplify specific regions of the identified novel variants from WES. Sanger sequencing was done using Applied Biosystems 3500 Genetic Analyzer at Bharathiar University, Coimbatore. The "Interactome network" mapping of novel candidate genes was done using Network Analyst (https://www.networkanalyst.ca/) to analyze the gene-gene interactions and to understand the functional relationships and molecular interactions among the candidate genes. The KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway Database (https: //www.genome.jp/ kegg/pathway.html) provides a collection of metabolic pathways, signalling pathways, and other functional modules.

The biological consequence of specific genetic alterations was validated in phase III, and their potential impact on breast cancer is assessed. Mutation profiling using Sanger sequencing in breast cancer involves the systematic analysis of genetic alterations within tumor DNA, revealing critical insights into the specific mutations and genomic changes that

drive breast cancer development and progression. Three new variants were detected in different patients: *MRPL13* (c.380T>C) in BC-2 patient, *MYBPC3* (c.2816G>A) in BC-4 patient, and *PTCH1* (c.1889G>A) in BC-4 and BC-6 patients. *MRPL13* (Mitochondrial Ribosomal Protein L13) is integral to mitochondrial ribosomal function, potentially impacting energy production and contributing to various diseases, including cancer. *MYBPC3* (Myosin Binding Protein C), primarily expressed in cardiac and skeletal muscles, is typically linked to cardiomyopathies. *PTCH1* (Patched 1), a crucial player in the hedgehog signalling pathway, is implicated in various cancers, such as basal cell carcinoma and medulloblastoma. Initially identified via exome sequencing, these variants were subsequently validated using Sanger sequencing. This approach enables the identification of potential therapeutic targets and the development of personalized treatment strategies for breast cancer patients.

The touchdown PCR successfully amplified the desired DNA fragments corresponding to the target region of interest. Subsequent gel electrophoresis of the PCR product, compared to the ladder DNA, confirmed that the size of the amplified DNA fragment closely matched the anticipated amplicon size as designed for the primer. This alignment between the amplicon and expected size strongly indicates that the PCR product accurately represents the intended DNA region and is suitable for further Sanger sequencing analysis.

Sanger sequencing was performed for the newly identified variants in breast cancer patients. It was determined that the *MRPL13*: c.380T>C (p.Leu127pro) variant in BC-2 patient and *PTCH1*: c.1889G>A (p.Arg630His) variant in BC-6 patient was not validated through sequencing, implying a low frequency or absence of variants in the patients. In contrast, the successful validation of *MYBPC3*: c.2816G>A (p.Arg939Gln) and *PTCH1*: c.1889G>A (p.Arg630His) mutations in the BC-4 patient through Sanger sequencing confirms the presence of these specific mutations in the tumor DNA. The *MYBPC3*: c.2816G>A variant in BC-4 patient verified its presence with a nucleotide change from G to A at position 2816 and an amino acid change from R to Q (Arg to Gln) at position 939. This validation occurred within the forward primer binding site at 131bp and the reverse primer binding site at 161bp. Additionally, the *PTCH1*: c.1889G>A (p.Arg630His) variant in BC-4 patient was also validated, indicating its presence with nucleotide change from G to A at position 1889 and amino acid change from R to H (Arg to His) at position 630. The variant forward primer is situated at the 129bp, while the reverse primer binding site is located at the 152bp.

Overall, mutations within the *MYBPC3* gene can give rise to a range of cardiac conditions, primarily hypertrophic cardiomyopathy (HCM). The *PTCH1* gene encodes the

Patched homolog 1 protein, a critical Hedgehog (Hh) signalling pathway component. This pathway is highly conserved and pivotal in normal embryonic development. The successful validation of *MYBPC3* and *PTCH1* variants in our study has precisely identified and characterized genetic mutations linked to diverse conditions, encompassing cardiac health and potential cancer susceptibility. We are the first to confirm the presence of cardiomyopathy-related and tumor-suppressor genes within the primary breast tumor sample. Our findings mark a significant advancement in the field, paving the way for comprehensive and personalized approaches to the diagnosis and treatment of breast cancer patients with concurrent cardiomyopathy-related genetic alterations.

The "Interactome network" mapping of candidate genes aimed to identify and catalog the interactions among various biomolecules, while KEGG pathway enrichment analysis was conducted on the genes within the network. Specifically, the *MYBPC3* gene interacted with 10 genes: *ALK, FBXO32, NCF1, BLNK, CUL1, CAPN1, TNNI3K, TTN*, *TRIM63*, and *SMURF2*. Conversely, the *PTCH1* gene exhibited interactions with 16 genes: *GLI1, SMO, CCNB1, DNAJA3, SMURF2, IHH, HHIP, SHH, GRK2, SMURF1, SUFU, CDON, GPC3, CDK1, GLI3*, and *GLI2*. The mutations in *MYBPC3* and *PTCH1* can lead to mutations in these interacted genes and disrupt their normal function. These genetic alterations may contribute to disease progression because the interacted genes involve various cellular processes and signalling pathways. The *CUL1* gene highly interacts with *MYBPC3*, and this interaction is implicated in various cellular pathways, including the hedgehog signalling pathway, TGF-beta signalling pathway, ubiquitin-mediated proteolysis, protein processing in the endoplasmic reticulum, and circadian rhythm.

In the first phase of our research (epidemiological study), it was observed that patients with breast cancer experience disturbed sleep cycles, leading to circadian rhythm disruption. The second phase of the study (exome analysis), showed all breast cancer patients were prone to cardiomyopathy, which may also be attributed to circadian rhythm disruption. Understanding the presence of these mutations is pivotal for gaining insights into the mechanisms underlying these health conditions and potentially identifying novel approaches for prevention and treatment.

Phase IV

Understanding the mechanisms of cell proliferation in breast cancer is essential for the development of targeted therapies. Through sequencing and bioinformatics studies, we have identified mutations that disrupt the standard regulatory mechanisms of cell growth and division. Elevated mitotic activity in tumors is linked to a poorer prognosis. So, in the fourth

phase of our research, we employed artificial intelligence to detect mitotic cells in histopathology images of breast cancer. Manual scoring of mitosis is time-intensive and laborious. In our study, we implemented AI-based mitotic detection to address this challenge, improve diagnostic accuracy, and reduce the overall time investment. We are the first to identify mitosis in real-time breast cancer histopathology images.

A total of 267 breast cancer histopathology images were collected from the ICPR 2014 dataset (MITOS-ATYPIA-14 grand challenge) dataset for the study. The real-time histopathology images were collected from the six breast cancer patients whose tumor DNA genetic alterations were studied by whole exome analysis. A total of 31 real-time breast cancer histopathology images were utilized for the study. The proposed technique was implemented using a sci-kit-learn machine learning package on a desktop computer with a Windows 10 operating system. The dataset was randomly divided into two sets: 80% for training and 20% for testing/validation. The architecture of the CNN consists of input, convolution, and output layers.

The total number of parameters trained is 7,401,697 were calculated based on the weights and biases of the model's layers. The Adam optimizer was used with a learning rate of 0.001. Adam was an optimization algorithm that adapted learning rates per parameter, leading to efficient training with 'binary_crossentropy' for loss function to detect mitosis. The model was trained for 25 epochs (complete passes through the training dataset) using the provided training data (x_train and y_train). Based on these criteria, the accuracy score was computed by dividing the number of accurate predictions by the total number of predictions made.

An integrated framework incorporating a more extensive set of histopathology images was constructed using machine learning algorithms. The system utilizes deep learning algorithms, demonstrating remarkable performance in medical image analysis. This framework provides both pathologists and clinicians with an efficient and accurate visualization of breast cancer progression and facilitates informed treatment decisions in a time-effective manner.

The model achieved a high accuracy of 0.96, which suggested that the model had successfully learned to recognize and generalize patterns present within the training data. The selected CNN architecture, featuring convolutional layers and max-pooling, proved effective in learning hierarchical features from the images. Integrating a dropout layer prevented overfitting by reducing the risk of neuron co-adaptation during training. The ample training data enabled the model to acquire diverse features and generalize effectively to various

samples. Using the Adam optimizer with an appropriately set learning rate facilitated efficient model convergence. Additionally, two plots were generated to analyze the progress of training.

The developed code loads and pre-processes histopathology images and categorizes them based on mitosis scores. The constructed sequential CNN model for mitosis detection effectively trains the model and analyses the training process using loss and accuracy graphs. The developed CNN model can be used for the detection of mitosis in the medical field for breast cancer diagnosis and prognosis.

Deep learning techniques have greatly improved the accuracy and efficiency of mitotic cell detection in breast cancer histopathological images. Various approaches have been explored, combining deep convolutional neural networks (CNNs) with innovative methods to tackle the challenges posed by mitotic cell detection. The proposed results showed consistent improvement and higher accuracy compared to current methods employed. Integrating deep learning techniques demonstrates the potential to enhance breast cancer prognosis and assist pathologists and researchers in histopathological analysis.

Conclusion

Integrating epidemiological, genetic, and image analysis approaches has opened doors for innovative prevention and treatment strategies. Epidemiology study helps in understanding the impact of social determinants of health on breast cancer diagnosis, treatment outcomes, and healthcare disparities. Characterizing the landscape of genetic alterations helps identify mutations, variations, and genetic signatures associated with the disease. The discovery of novel variants in *MYBPC3* (c.2816G>A) and *PTCH1* (c.1889G>A) involved in breast cancer development and cardiomyopathy was identified, which led to the identification of potential therapeutic targets and personalized treatment approaches based on genetic alterations. Automated quantification of the mitotic score in histopathology images provides a standardized approach to assessing breast cancer progression. The multifaceted understanding of breast cancer emphasizes the importance of personalized risk assessment, comprehensive care, and early detection in the battle against breast cancer, offering hope for improved patient outcomes and quality of life.

Bibliography

1. Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., and Soerjomataram, I. (2022) Current and future burden of breast cancer, Global statistics for 2020 and 2040. The Breast., 66, 15-23. doi: 10.1016/j.breast.2022.08.010
2. Campos, F. A. B., Rouleau, E., Torrezan, G. T., Carraro, D. M., Casali da Rocha, J. C., Mantovani, H. K., and Santana dos Santos, E. (2021) Genetic landscape of male breast cancer. Cancers., 13, 3535. doi: 10.3390/cancers13143535
3. Conway, J. R., Warner, J. L., Rubinstein, W. S., and Miller, R. S. (2019) Next-generation sequencing and the clinical oncology workflow, data challenges, proposed solutions, and a call to action. JCO Precis. Oncol., 3, 1-10. doi: 10.1200/PO.19.00232
4. Glaviano, A., Foo, A. S., Lam, H. Y., Yap, K. C., Jacot, W., Jones, R. H., and Kumar, A. P. (2023) PI3K/AKT/mTORsignalling transduction pathway and targeted therapies in cancer. Mol. Cancer., 22, 1-37. doi: 10.1186/s12943-023-01827-6
5. Huang, J., Zhang, L., Wan, D., Zhou, L., Zheng, S., Lin, S., and Qiao, Y. (2021) Extracellular matrix and its therapeutic potential for cancer treatment. Signal. Transduct. Target. Ther., 6, 153. doi: 10.1038/s41392-021-00544-0
6. Iacoviello, L., Bonaccio, M., de Gaetano, G., and Donati, M. B. (2021) Epidemiology of breast cancer, a paradigm of the "common soil" hypothesis. Semin. Cancer Biol., 72, 4-10. https://doi.org/10.1016/j.semcancer.2020.02.010
7. Ibrahim, A., Lashen, A., Toss, M., Mihai, R., and Rakha, E. (2022) Assessment of mitotic activity in breast cancer: revisited in the digital pathology era. J. Clin. Pathol., 75, 365-372. doi: 10.1136/jclinpath-2021-207742
8. Jeibouei, S., Akbari, M. E., Kalbasi, A., Aref, A. R., Ajoudanian, M., Rezvani, A., annd Zali, H. (2019) Personalized medicine in breast cancer: pharmacogenomics approaches. Pharmacogenomics Pers. Med., 59-73. doi: 10.2147/PGPM.S167886
9. Osei-Afriyie, S., Addae, A. K., Oppong, S., Amu, H., Ampofo, E., and Osei, E. (2021) Breast cancer awareness, risk factors and screening practices among future health professionals in Ghana: A cross-sectional study. PLoS One., 16, e0253373. doi: 10.1371/journal.pone.0253373
10. Pan, X., Lu, Y., Lan, R., Liu, Z., Qin, Z., Wang, H., and Liu, Z. (2021) Mitosis detection techniques in H&E stained breast cancer pathological images: A comprehensive review. Comput. Electr. Eng., 91, 107038. https://doi.org/10.1016/j.compeleceng.2021.107038

11. Pantanowitz, L., Hartman, D., Qi, Y., Cho, E. Y., Suh, B., Paeng, K., and Cho, S. Y. (2020) Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses. Diagn. Pathol., 15, 1-10. doi: 10.1186/s13000-020-00995-z
12. Qin, D. (2019) Next-generation sequencing and its clinical application. Cancer Biol. Med., 16, 4-10. doi: 10.20892/j.issn.2095-3941.2018.0055
13. Rajaneesh, N. (2021) Risk Factors in Breast Cancer: Work on Preventing Recurrence. Notion Press., 1-23.
14. Sahu, R., and Pattanayak, S. P. (2020) Strategic developments and future perspective on gene therapy for breast cancer: Role of mTOR and Brk/PTK6 as molecular targets. Curr. Gene Ther., 20, 237-258. doi: 10.2174/1566523220999200731002408
15. Sathishkumar, K., Chaturvedi, M., Das, P., Stephen, S., and Mathur, P. (2022) Cancer incidence estimates for 2022 & projection for 2025: Result from national cancer Registry Programme, India. Indian J. Med. Res., 156, 598-607. doi: 10.4103/ijmr.ijmr_1821_22
16. Taha, Z., and Eltom, S. E. (2018) The role of diet and lifestyle in women with breast cancer: An update review of related research in the Middle East. Biores. Open Access., 7, 73-80. doi: 10.1089/biores.2018.0004

Publications

- **Suganya, K**., Sudha, B., Poornima, A., Senthil Kumar, N., and Sumathi, S. (2022) Reduced Expression of *SFRP1* is Associated with Poor Prognosis and Promotes Cell Proliferation in Breast Cancer: An Integrated Bioinformatics Approach. Indian J. Gynecol. Oncol., 20, 4. https://doi.org/10.1007/s40944-022-00650-z
- **Suganya, K**., and Sumathi, S. (2022) A Comprehensive Review on Role of Nutrition in Management of Breast Cancer. Indian. J. Nutr. Diet., 59, 506-523. https://doi.org/10.21048/IJND.2022.59.4.30271
- **Suganya, K**., Sumathi, S., Karthikesh, K., Bhargavi, S., and Sethumadhavan, T. (2023) Performance Analysis of Various Filters for Denoising Breast Cancer Histopathology Images. Indian J. Gynecol. Oncol., 21, 4. https://doi.org/10.1007/s40944-023-00761-1